基于深度神经网络的中文新闻文本分类方法

郑创伟 王 泳 邢谷涛 谢志成 陈义飞
（深圳市创意智慧港科技有限责任公司，广东 深圳 518034）

摘要:【目的】文章比较多个基于深度神经网络的中文新闻文本分类模型，旨在找到准确度较高的方法用以实际工作，为中文新闻文本分类提供更加高效的方法。【方法】对文本分类技术和中文新闻分类进行了梳理和归纳，对中文新闻文本的特征和预处理进行了阐述，详细介绍 FastText 算法、Bert 分类算法、TextCNN 算法和 TextRNN 算法。【结果】四种深度神经网络算法均可以应用于中文新闻文本分类，可以有效处理信息紊乱问题以及快速准确进行分类。【结论】通过对四种深度神经网络算法进行试验和效果对比，发现 FastText 模型在实际工作中的文本分类效果最为优异。

关键词：深度神经网络；文本分类；中文新闻；自然语言处理 中图分类号：TP183 文献标识码：A
文章编号：1671–0134（2023）03–147–05 DOI：10.19483/j.cnki.11–4653/n.2023.03.033
本文著录格式：郑创伟，王泳，邢谷涛，谢志成，陈义飞．基于深度神经网络的中文新闻文本分类方法 [J]. 中国传媒科技，2023（03）：147–151.

导语

随着信息时代的高速发展，网络信息呈现爆炸式增长。新浪、今日头条等一些主流新闻网站，每天提供数以百万计的新闻数据，然而这些爆炸式增长的数据给网站带来了巨大的挑战。新闻文本分类可以有效地对文本进行快速准确分类，提高网站的工作效率，成为近些年来的研究热点。新闻文本分类属于文本分类的一个子任务。文本分类广泛应用于各个领域，如网页分类、微博情感分析、用户评论挖掘等，是自然语言处理中使用率最广泛的技术之一。文本分类最重要的作用是可以有效处理信息紊乱问题，尤其是对海量信息而言，更能够帮助用户快速、高效准确地定位所需信息，从而更加高效地分析数据。[1]

本文对新闻文本分类技术进行探究和阐述，主要包括分类特点等，并通过实验指出各个算法的优劣所在，预测未来新闻分类的发展趋势。

1. 相关研究

1.1 中文新闻分类概述

中文文本是一种无法被计算机处理的非结构化数据，要转化为结构化数据。结构化数据的过程首先要进行数据预处理，然后用一些特征提取的方法就可以使用。[2] 特征提取可以概括为以下三类:（1）词袋模型。（2）特性权重计算。（3）向量空间模型。词袋模型指忽略词序和语法，将文本仅仅看作是一个词集合。若词集合共有 N 个词，每个文本表示为一个 N 维向量，元素为 0/1，表示该文本是否包含对应的词。特性权重计算一般有布尔权重、TFIDF 型权重，以及基于熵概念权重等几种方式。向量空间模型指以词袋模型为基础，通过特征选择来降低模型维度，并且利用特征权重来进行二次计算。[3] 通过上述方法，可以将非结构化的文本转化为结构化的数组，从而进行文本分类。

基于传统的机器学习方法，主要可以概括为特征工程 + 浅层分类模型。基于机器学习分类方法中，会将数据集按照一定比例分为训练集和测试集，然后通过不断训练调整分类模型的参数来达到更高的准确率，再利用测试集对该分类模型的分类效果进行评估。[4] 在分类过程中，可以利用相似语料对提取出的文本信息进行扩展，进而得到特征向量，或者利用支持向量机，以及信息增益的计算方式来选择特征，提高分类准确率。此外，还能够对词向量进行加权处理，这样能更加精准区分不同词条的重要程度，提高分类文本的准确率和效率。由于不同的任务对特征的要求不一样，所以具体问题需要具体分析。其中最主要涉及的技术为构建分类器，这是一种基于统计分类的方法，包括 SVM 和朴素贝叶斯分类算法等。[5]

基于深度学习的文本分类方法，利用 CNN/RNN 等网络结构自动获取特征表达，然后进行分类，从而端到端的解决问题。基于深度学习分类方法中，由于计算机性能不断提升，使得图像识别、自然语言处理等领域得到了快速发展。这种算法模拟了人的大脑中

神经元的连接与计算，在其神经网络中，一般包含输入层、隐藏层和输出层。层与层之间通过反向传播算法等对数据进行训练和计算，得到相应的训练模型。深度学习的方式，往往也意味着其隐藏层较多，每层负责学习的特征有所区别，最终将这些特征汇总在一起，完成更加精准的学习任务。[6]在对文本分类过程中，可以从用户特征信息、文本主题信息，以及评论关键词等角度出发，提取结构化文本中的特征信息，这样能够取得更好的分类效果。

2. 中文新闻文本分类研究

2.1 中文新闻文本特征

从文本分类的角度分析，中文新闻具有以下两个特征：（1）新闻需要文本分类。随着信息时代数据量爆发式增长，新闻也呈现指数型增长，如何从这些海量的数据当中获取需要的新闻成为一个热点问题。（2）新闻分类具有可行性。由于新闻数据的公开性，网络上充斥着大量的训练和测试数据。与此同时，随着分类算法快速发展，分类性能也越来越高。

2.2 中文新闻文本预处理

中文新闻的文本预处理主要是针对一些无实际意义的词进行识别和剔除，例如大量的停用词或噪声等，从而能够降低其对预处理的影响程度。[7]文本预处理的过程主要包括：分词、降噪、词性标注、剔除停用词等。

2.2.1 分词

在中文新闻分词过程中，没有类似英文中间空格的断开分词特征，因此就需要对其进行更多处理，例如，使用向前向后最大匹配算法等，可以使用基于字典或者基于统计的方法进行分词。中文分词主要是解决中文文本中缺少形式上的分隔符这一难题，中文分词所使用的技术主要有以下几种：第一，基于字符串匹配技术，这种方法的关键是必须建立统一的词典表，当句子开始进行分词时，先将句子进行拆分，拆分后再和之前建立的词典表进行匹配对比。第二，基于理解的分词方法，这种方法是让计算机通过神经网络算法去模拟人对句子进行理解和表达，进而可以识别中文词语，但因中文词语的语义较广，因此难度较大。第三，基于统计的分词技术，这种方法的最基本思维就是利用了统计学和概率等，认为分词是一个概率最大化问题，基于所构建的语料库，统计相邻的字组成的词语出现的概率，按照概率值进行分词。

2.2.2 降噪

对中文新闻信息的降噪，主要是去除网页上杂乱的文字和图片，只保留经过工整排版的正文部分。如果遇到短文本，还需要剔除一些表情符号、转发关系等，仅保留纯文本用于后续分析和处理。在降噪过程中可能涉及特征抽取或特征降维这一操作，其可以有效降低算法计算的开销、去除噪声，能够提升模型的训练速度。

2.2.3 词性标注

降噪完成后，需要对中文新闻中的词语进行词性标注，包括名词、动词、形容词、副词等。词性标注的作用主要体现在后续对文本进行识别和分类的过程中，经过词性标注后，处理效率能够大大提升。

2.2.4 停用词或无意义词过滤

第一种方法是根据已制定的停用词表进行处理，停用词表中一般包含语气词、标点符号等，在对新闻信息分词去噪后，对其进行遍历，遇到与停用词表中相同词语时，将其剔除。这种方法可控性较好，效率较高，能够随时对停用词表进行修改。第二种方法是计算语料库中词语出现的频率，然后选择出现频率较低或次数较少的词语进行剔除。但这种方法计算量较大，会消耗较多资源，有时还可能将某个出现频率较低但影响较大的词语误删除。

2.3 中文新闻文本分类的主要模型方法

文本分类是根据文本语义内容来对其进行归类的一个过程，文本数据集合与类别集合之间可以用3–1函数表示：

$$(d_i,\ c_j) \rightarrow \Phi(d_i,\ c_j) = \begin{cases} T,\ if\ \ d_i \in c_j \\ F,\ if\ \ d_i \notin c_j \end{cases} \quad （式 3–1）$$

基于上文提到的新闻特征，将文本分类应用到新闻领域有重要实际意义。新闻文本分类具有以下三个特点[8]：（1）文本分析要考虑标题的重要性：新闻标题是对一篇文章的高度概括，它对新闻的分类有很大的辅助作用；（2）文本表示要考虑新闻特征：充分分析新闻文本的特性，进而优化文本表示方法，有助于提高网络新闻的分类效果；（3）分类标准偏向主题而非学科。因此，本研究针对实际工作中遇到的新闻数据，基于深度学习的分类算法，采用了FastText、TextCNN、BERT、TextRNN等模型进行计算和训练。在训练过程中要注意对数据集进行分类，预设的判断条件要尽可能科学，例如，考虑用梯度下降的反向传播算法来更新权值，从而使得准确率逐步提高，达到更好的训练效果。

2.3.1 FastText 模型

FastText模型主要包括输入层、隐含层和输出层（如图1），与大型神经网络结构相比其较为简单，运行

效率较高，它在保证分类准确率的同时，还能够进一步提升训练速度。[9] 在输入层中，将文本当作一个由词构成的集合，生成表征文本的向量，在此过程中的关键操作为对文本中出现的词实施叠加平均操作，最后利用该向量完成多分类任务。此算法的优点还体现在可以无须进行预训练步骤,其可以自发训练词向量，将单词序列作为输入，并且使用层次 softmax 函数对分类进行加速，以及预测这些类别的概率分布。这种以霍夫曼编码树形式来建立层次的方法，大大降低了计算复杂度。

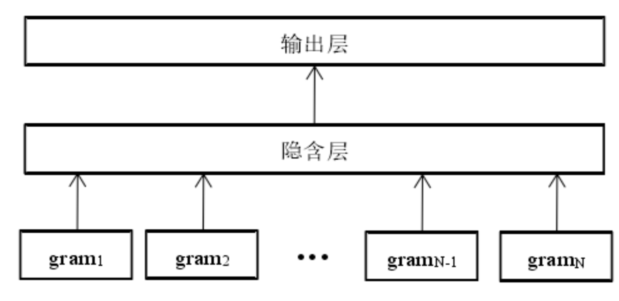


图 1　FastText 模型结构

2.3.2　TextCNN 模型

选择合适的中文文本分类算法是中文文本分类的核心，这需要对每种算法有一定程度的了解，同时要对新闻文本分类任务有清晰的认知。使用 TextCNN 处理文本并进行分类，就必须对信息进行数据预处理操作，以便后期达到更好的分析效果，具体包括向量化、词向量初始化等。在文本分类中，TextCNN 模型应用最为广泛，尤其在工业领域应用更为成熟，已经取得了较为优异的输出效果，其网络结构较为简单，因此模型可以使用较少的参数进行训练，有效节约计算开支，提高了训练速度。CNN 主要运用在图片分类领域，而 TextCNN 则是其一种变形，能够用于文本分类，结构示意图如图 2 所示，词向量经过不同卷积核运算后得到对应的特征向量,再经过池化层后得到全连接层，此时映射运算就能够将高维数据转换为低维数据。[10] TextCNN 的可解释性较弱，需要人工对其进行指导干预，对卷积核的尺寸进行设定，并且需要对模型进行手工调优。TextCNN 模型通过利用一个 k 维向量来代表某句子中的一个单词，这些单词会做成一个词典以供文本输入后使用。文本输入后会将每个单词对应一个一维向量，最终将一整个句子转换成一个二维矩阵后卷积,而此时卷积核的列维度就与输入的维度相同，并且卷积核的大小可以根据实际情况进行调整，滑动步长的范围往往控制在 2 ~ 5 个单词之间。

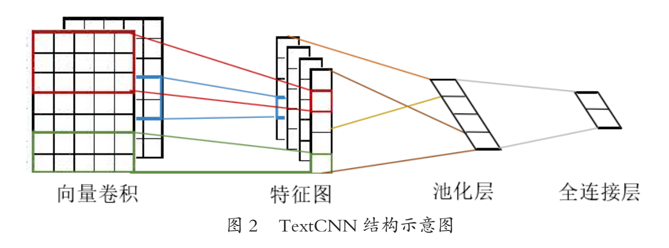


图 2　TextCNN 结构示意图

从图 3 TextCNN 算法流程图中能够看出，在输入文本信息后，开始对文本进行数据预处理，此时使用到词嵌入、词向量初始化、向量维度变换等方法。数据预处理完毕后，使用 Text CNN 进行训练，通过卷积、最大池化、Softmax 方式输出分类结果。最后对输出的损失值进行判断，如果超过了设定的阈值，则以梯度下降的反向传播算法进行循环更新，直到小于或等于设定的阈值则训练结束。常用的梯度下降方法为批量梯度下降法,即在每一次迭代过程中都需要更新梯度。梯度下降的优点在于其利用矩阵计算所有样本数据，可对数据进行并行处理；缺点在于当数据量较大时，每次计算所有数据会使得训练效率有所降低。

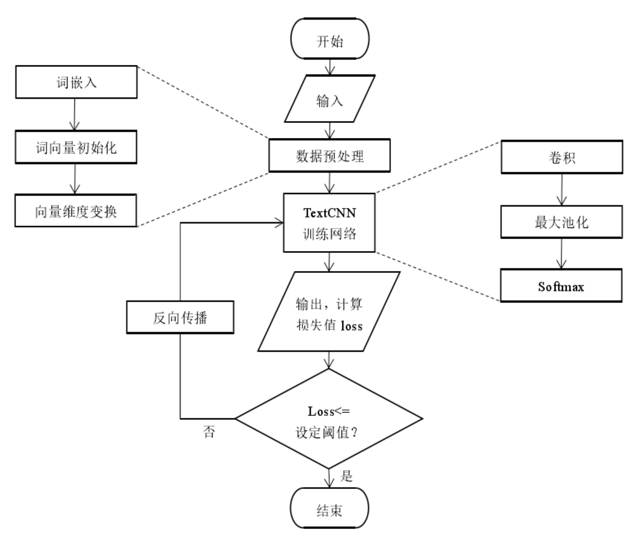


图 3　TextCNN 算法流程图

2.3.3　Bert 模型

Bert 最早是谷歌团队发明的一种语言模型，它由多个 Transformer 的 Encoder 叠加而成，模型结构如图 4 所示。Transformer 结构是采用一种注意力机制，在读取数据信息时会一次性读取文本序列，不仅能够提高读取效率，还能够更方便的基于单词的上下文进行语义学习，增强了对上下文语义的理解，也与中文语言表述更接近。这种方法对新闻文本分类而言，可以解决数据稀疏、上下文依赖性过高等难点，使得文本分类性能更加高效，满足更加精准性的需求。

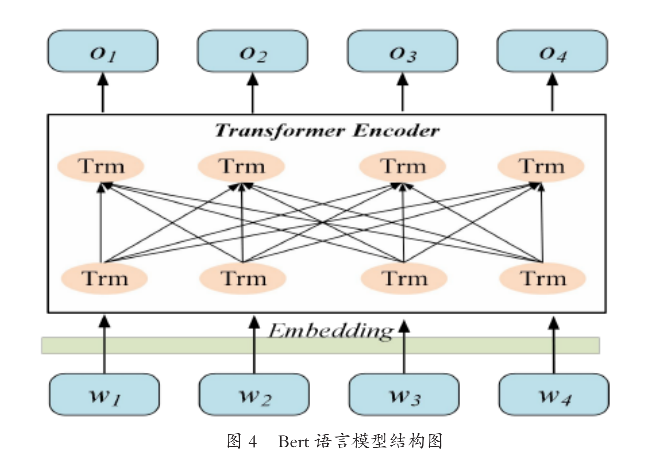


图 4 Bert 语言模型结构图

该模型的输入层主要是利用 Bert 模型算法进行预训练，进而能够以文本语义向量表示。在句子开头和结尾处需要进行标记，然后对读取到的数据进行处理，采用映射索引的方法对文字和标签进行切分，然后将每一个词嵌入转换为一维语义向量。再通过 Transformer Encoder 堆叠，完成双向语义特征学习及向量表示。在特征抽取层，要通过 Bert 模型进行进一步微调，结合注意力机制对文本特征进行提取，通过这种机制能够更加聚焦于数据内部的相关性，利用词向量加权的方式提高模型运算效率。[11]Bert 算法模型就是由多个 Transformer 的 Encoder 部分叠加的深层次网络，该方式一次性读取整个文本序列，因此可以用于对某个单词上下文语义进行学习，增强了对上下文语义学习的理解能力，在一定程度上更加接近人类语言。同时还会对文本进行特征抽取，示意图如图 5，具有全局时序最优等特征，可以提取文本信息中上下文语义信息，具体实现过程中需要利用 Tensorflow 库函数来搭建双向网络操作函数。在输出层，主要是对每个样本所属的标签做概率预测，对文本信息能够进行高效提取，然后通过全连接的方式提高分词准确率。这种全连接方式利用了激活函数和数据线性变换的方式来提高计算效率，并且采用梯度下降算法来进行参数学习和 Dropout 策略防止模型过拟合问题。

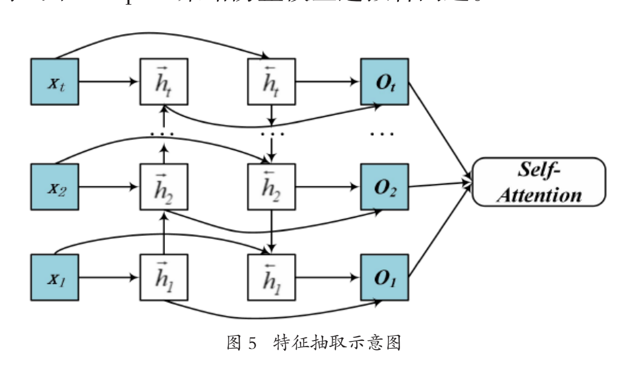


图 5 特征抽取示意图

2.3.4 TextRNN 模型

该递归神经网络模型又名文本循环神经网络，利用该模型在中文新闻文本分类时，能够捕获更长的序列信息，它避免了 CNN 算法中不能延展序列长度的缺陷，并且在进行参数调节时较为简单，可以更加准确地表达上下文信息。在 RNN 算法中，输出的结果并不仅仅是由矩阵和卷积计算得到的，其会根据计算得出一个 State，并且会持续影响后续的计算，这样经过 N 个样本的输出，就能够使得结果具备一定的序特征。这就使得输入数据的状态可以在自身神经网络中进行循环处理，并且产生时间关联。TextRNN 模型的特别之处在于其同一隐藏层的节点之间是存在连接的，并且将时间关系作为影响数据间关系的变量，它不仅考虑当前的输入，还赋予网络对过去的记忆。在其隐藏层中，数据可能会从第一个隐藏层中输出后，再加上一定的权重进入第二个隐藏层，也就是说在向下一层输入时，会将某一时刻的隐藏状态神经元和这一时刻的文本特征一起输入。最后经过的不断循环和递归，再反向调整各层的连接权重，得到最优化参数。但正是由于这种结构，使得 TextRNN 后一个时刻的输出会依赖前一个时刻的输出，因此无法并行处理，降低了训练效率。[12]

从图 6 TextRNN 网络结构中可以看出，数据按时间序列展开后，能够得到一个 T 维向量，U 为输入层到隐藏层的权重，权重越大则代表输入信息量越多。横向 W 则代表前一个隐藏层到后一个隐藏层的权重，V 则代表从隐藏层到输出层的权重。要注意的是，RNN 在处理序列信息时，有时会偏向最后输入的信息，这就可能导致早期信息丢失的问题，因此在初始化权重时，要尽可能避免极大或极小值，并且加入 LSTM（长短期记忆网络）和 GRU（门控循环单元）。

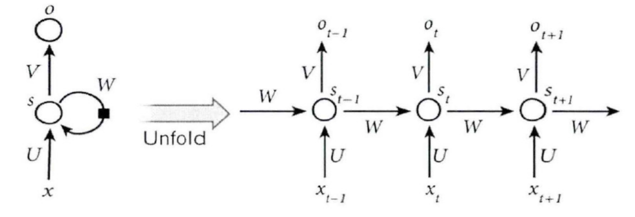


图 6 TextRNN 网络结构

2.4 中文新闻文本分类实验

2.4.1 数据集介绍

笔者提供了一个新闻和公司相关的数据集，数据集是通过对某网的金融数据进行筛选过滤生成，包含 40 万篇新闻，都是经过预处理过后的文本，均为 UTF-8 纯文本。在原始网站的基础之上，将数据集划分出 1000 个类，每一个类代表一家公司。将用一些主

流的分类算法测试模型的性能。

2.4.2 实验结果

实验需要对测试数据集分类结果的准确性进行评价，若结果不在合理范围内，则需要返回特征选择阶段再次完成特征选取，重复这一步骤直到结果处于合理范围内。其分类标准主要包括准确率和召回率，准确率可以表示文本分类模型的准确程度，但仅准确率高而召回率很低，则代表没有把本应预测出来的标签类别预测出来，尤其是对于非均衡样本，有时会把小类样本预测成为大类样本；或者某个多标签分类模型，可能会出现特征和模型过拟合的现象，这也会导致召回率较低，因此在实验时要加以注意。

笔者分别使用FastText算法、Bert分类算法、TextCNN算法、TextRNN算法来分别在数据集上进行测试，评估标准准确率和召回率，实验结果如下表所示：

分类算法	准确率（%）	召回率（%）
FastText	88.5	88.5
Bert	85	84
TextCNN	87	88
TextRNN	86.5	86

同时本研究还在THUCNews上测试了几种方法的准确度和召回率，实验结果如下表所示：

分类算法	准确率（%）	召回率（%）
FastText	86.5	86.5
Bert	83.4	81.2
TextCNN	87.9	88.5
TextRNN	86.7	86.5

3. 总结

本文在对中文文本分类进行梳理和研究的基础上，认为以下几个方向将成为研究的热点：（1）基于无监督学习模式的新闻文本分类：网络上充斥着大量无监督的数据，如何利用好这些数据，将成为一个热门研究；（2）多层次新闻文本分类：充分利用分类体系的层次信息，采用逐层分类思想进行多层次文本分类，能有效地降低分类算法的复杂度，同时保证分类精度，值得进一步研究。（3）跨模态的新闻文本分类：新闻文本分类主要考虑文本信息，新闻中一些其他模态的信息被忽略，如何利用这些信息辅助分类，充分融合好文本信息和图片信息，也是一个研究热点。同时，本研究讨论了新闻文本分类等相关研究，分别介绍了FastText模型、TextCNN模型、BERT模型以及TextRNN模型。经过实验，FastText模型在实际工作中的文本分类效果最为优异，而TextCNN模型在THUCNews上的文本分类最为优异。

参考文献

[1] 李泽魁，孙霏，陈珺．新闻媒体领域中文语义分析技术智能化、知识化之路的研究与探索 [J]. 中国传媒科技，2018（8）：35-37.

[2]Li Z，Shang W，Yan M．News text classification model based on topic model[C]// IEEE/ACIS International Conference on Computer & Information Science. IEEE，2016.

[3] 李可悦，陈轶，牛少彰．基于BERT的社交电商文本分类算法 [J]. 计算机科学，2021（2）：87-92.

[4] 贾澎涛，孙炜．基于深度学习的文本分类综述 [J]. 计算机与现代化，2021（7）：29-37.

[5] 谭辛．政策解读大数据分析应用的实践探究 [J]. 中国传媒科技，2019（3）：22-23.

[6] 刘萌．人工智能技术在媒体融合中的运用研究 [J]. 中国传媒科技，2021（11）：154-156.

[7] 李泽魁，孙霏，陈珺．新闻媒体领域中文语义分析技术智能化、知识化之路的研究与探索 [J]. 中国传媒科技，2018（8）：35-37.

[8] 贾红雨，王宇涵，丛日晴，林岩．结合自注意力机制的神经网络文本分类算法研究 [J]. 计算机应用与软件，2020（2）：200-206.

[9] 杨锐，陈伟，何涛，张敏，李蕊伶，岳芳．融合主题信息的卷积神经网络文本分类方法研究 [J]. 现代情报，2020（4）：42-49.

[10] 杜思佳，于海宁，张宏莉．基于深度学习的文本分类研究进展 [J]. 网络与信息安全学报，2020（4）：1-13.

[11] 郝超，裘杭萍，孙毅，张超然．多标签文本分类研究进展 [J]. 计算机工程与应用，2021（10）：48-56.

[12] 王迷莉．基于机器学习的文本分类研究 [J]. 科技创新与应用，2021（26）：70-72.

作者简介：郑创伟（1978-），男，广东汕头，高级工程师，研究方向为大数据、人工智能；王泳（1977-），女，湖南邵阳，中级工程师，研究方向为大数据；邢谷涛（1984-），男，海南文昌，中级工程师，研究方向为云计算；谢志成（1980-），男，广东汕头，中级工程师，研究方向为大数据、云计算；陈义飞（1981-），广东湛江，中级工程师，研究方向为大数据。

（责任编辑：张晓婧）